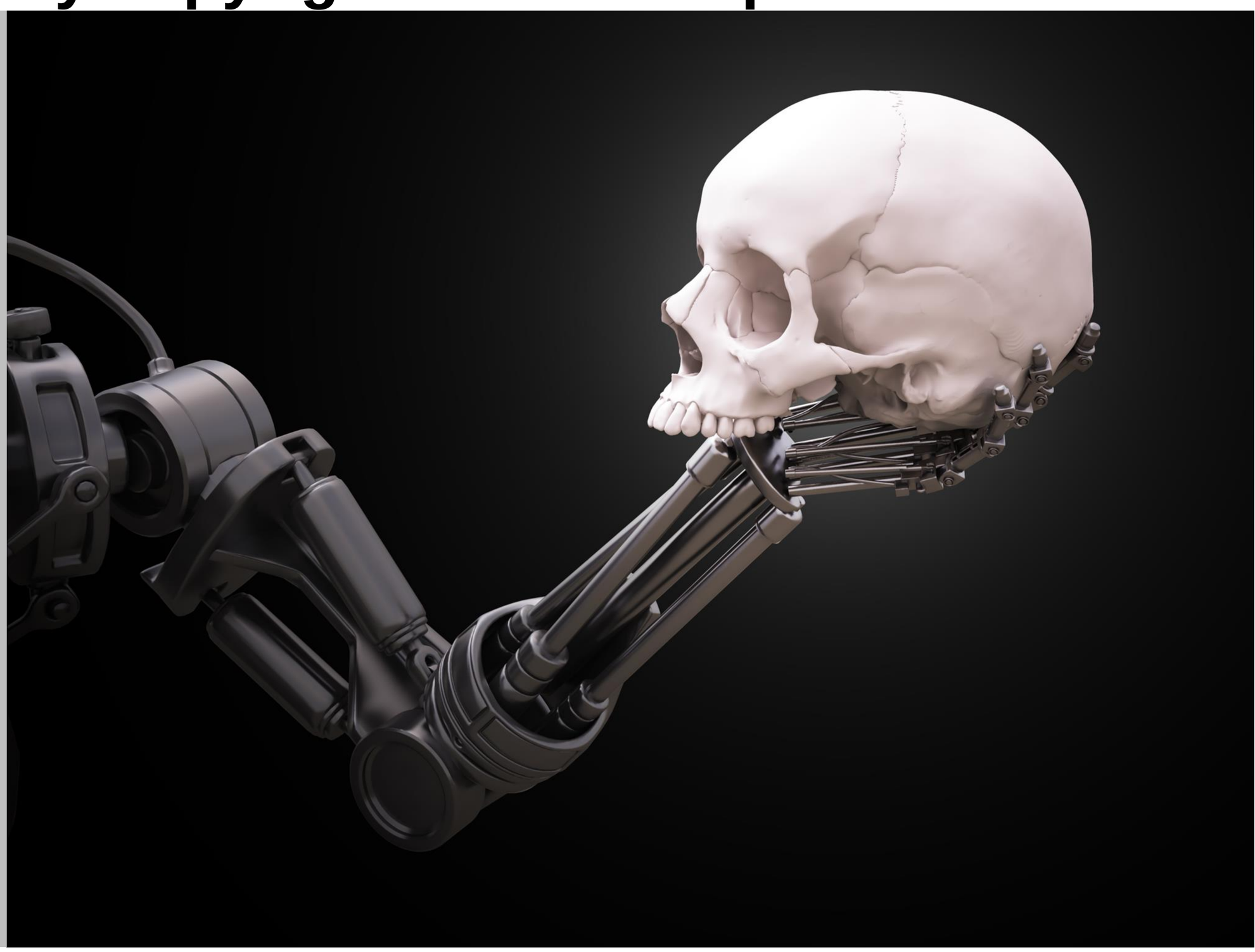
Christian Geib (2nd year PhD) Law christian.geib@strath.ac.uk

The Machines are Rising: TEXT MINING.

Why text mining is good news for research, the tax payer and why copyright law still impedes it.

Is the right to read the right to mine? Are researchers in danger of infringing copyright?



What is text mining?
Text mining denotes the automatic or semi-automatic analysis of large amounts of textual data by using computer programs. The aim is to discover unknown correlations and patterns to form new hypotheses.

Example: Swanson's early work[1]
Journal articles of two disparate disciplines were text mined. One set of articles reported blood changes that ameliorated the symptoms of Raynaud's disease. The other set of articles reported that fish oil caused the same blood changes. However, neither set had noticed the other set. The logical connection for a new treatment might have been overlooked had it not been for text mining.

The problem
Text mining involves scanning text and placing it into repositories. During that process at least one copy is made. This, if not permitted by author or publisher, is on its face infringing copyright. The threat of infringement could impede the adaption of this beneficial technology.

How does this relate to the tax payer?
Text mining enables researchers to obtain new insights from research that is already paid for, e.g. by text mining scientific journal articles and looking for hidden correlations. However, some journal publishers request extra licences for text mining even for journals that universities have already lawfully subscribed to, even after the taxpayer has already paid for the research that led to these journal articles in first place.

Solutions
Could text mining be covered by copyright exceptions such as 'fair dealing' or 'temporary copying'? If not, could new licensing models be the answer? If not, what could possible legislative responses be? What could researchers do to minimize the risk of infringement? Those are some of the questions this PhD thesis attempts to answer.

1 Swanson DR, 'Two Medical Literatures That Are Logically but Not Bibliographically Connected' (1987) 38 Journal Of The American Society For Information Science